AD_____

AWARD NUMBER DAMD17-94-J-4076

TITLE: Development of a Common Database for Digital Mammography Research

PRINCIPAL INVESTIGATOR: Robert M. Nishikawa, Ph.D.

CONTRACTING ORGANIZATION: The University of Chicago
Chicago, Illinois 60637

REPORT DATE: October 1998

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1998	Annual (15 Sep 97 - 14 Sep 98)

4. TITLE AND SUBTITLE
Development of a Common Database for Digital Mammography Research

5. FUNDING NUMBERS
DAMD17-94-J-4076

6. AUTHOR(S)
Robert M. Nishikawa, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The University of Chicago
Chicago, Illinois 60637

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The purpose of this infrastructure project is to develop a large database of digitized mammograms that will be distributed free of charge to researchers working in all aspects of digital mammography. This database will facilitate and promote rapid development in digital mammography research. The database will consist of 1000 cases subdivided into 5 categories, 4 containing different breast lesions -- masses, microcalcifications, architectural distortions, asymmetric densities (both benign and malignant) -- and one containing normal mammograms. The mammograms will be collected and digitized (0.05-mm pixel size) at two sites: the Universities of Chicago and North Carolina. The database will be stored at the two sites and will be available over internet, and by mail on CD, tape and magneto-optical disks. To date 448 cases have been digitized. Each case consists of index and previous exams (each having four standard views) and up to two special-view mammograms (e.g., magnification views). The computer systems for the database have been assembled and are connected to the network. The first release of 50 cases with clustered microcalcifications will be made immediately. This trial release will be followed by another 50 cases of microcalcifications and 100 cases with masses.

14. SUBJECT TERMS
Breast Cancer
digital mammography, database, information systems, computer-aided diagnosis, image analysis, image processing

15. NUMBER OF PAGES
9

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

_____ Where copyrighted material is quoted, permission has been obtained to use such material.

_____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

_____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

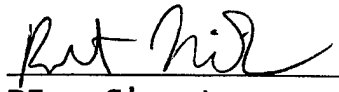
_____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

_____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

_____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

_____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

_____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.



10/14/98

PI - Signature Date

4. TABLE OF CONTENTS

5. INTRODUCTION

This research is to develop a large database of digitized mammograms that will be distributed free of charge to interested researchers. It is being funded by the USAMRMC as an infrastructure award and as such there it does not represent a research project per se. That is, there is no hypothesis that we are trying to prove. Therefore, this report is structured slightly different than a normal scientific research report -- heavy on the method and light on actual results. In this project, the procedure is the most important component, which is applied continuously in a straightforward manner to achieve the goal of creating the database of mammograms.

5.1 Nature of the Problem

In 1992, the National Cancer Institute identified digital mammography as an important area of research for reducing breast cancer mortality.[1] As a result, there has been a sharp increase in the number of researchers developing computerized methods for analyzing mammograms. This is due in part to the substantial potential benefit from developing an automated computerized system for assisting radiologists in interpreting mammograms. With a large number of investigators developing computerized analysis techniques, the likelihood of an accurate method being developed is high. Unfortunately, a major obstacle to rapid progress in developing a technique is that each investigator uses his or her own set of mammograms (database) to develop and evaluate the performance of his or her technique. As a result, it is not possible to compare the accuracy of different methods because the measured performance is dependent on the cases used for testing.[2] For example, by using "easy" cases for testing, a computer technique would apparently have a higher accuracy than if "hard" cases were used. A common database of mammograms that could be used by all investigators in the field would solve this problem.

5.2. Background: Previous work in the field

At a Biomedical Image Processing meeting held February 1993, in San Jose CA, 12 panelists discussed the design of a common database for research in mammographic image analysis.[3] Two of the panelists are investigators are on this proposal. Important considerations in developing the database are: (a) the cases selected, (b) the digitizer used, (c) organization of the database, (d) associated information to be included with images, (e) "truth" for each case, (f) format of image files, (g) distribution of the database, and (h) rules on using the database.

There have been several small databases released for general use. However, all have several limitations to due to insufficient spatial resolution, insufficient grey-scale resolution, and/or too small a number of cases. The database that we are developing will have none of these limitations. There is now underway the development of another mammographic database. This database differs in the one being developed in project because a smaller pixel size is being used and they are not including previous films as is being done in this project.

5.3. Purpose

The purpose of this proposal is to develop a database of digital mammograms that can be used by researchers who (1) are trying to determine the image quality requirements of detectors for digital mammography; (2) are developing image processing techniques to optimize the displayed digital mammogram; (3) are developing computerized methods for analyzing mammograms; (4) are studying the effects of image compression methods on image quality; (5) are developing methods for remote transmission of mammograms; and (6) are studying the relationship between image quality and diagnostic accuracy. This database also could be used as a resource for teaching radiology residents and for testing the performance levels of mammographers.

The specific aims of this proposal are:

1. Collect and digitize 200 cases in each of 5 different categories, mammograms exhibiting: (i) clustered microcalcifications, (ii) masses, (iii) architectural distortions, (iv) asymmetric densities, and (v) no lesions (i.e. normals).

2. Make these cases available to other researchers either over computer network (Internet) or by sending images on computer tape or CD. The database will be distributed as widely as possible so that comparisons of different computerized analysis techniques can be standardized.

5.4. Method of Approach

Task 1: Collect and digitize mammograms, Months 1-48. (See Figure 1.)

a. Retrieve from film library cases with pathologically-proven lesions (clustered microcalcifications, breast masses, architectural distortion, asymmetric densities), 100 cases of each type and 100 normals (cases without lesions) from each site [University of Chicago (UC) and University of North Carolina (UNC)] for a total of 1000 cases during the entire funding period.
b. At each site, digitize retrieved films and outline the location of the lesion in each abnormal image. The outline will be stored together with the images but in a separate file.
c. Send normal cases and asymmetric density cases that were digitized at UC to UNC; and send cases containing masses, microcalcifications, and architectural distortion that were digitized at UNC to UC.
d. Selectively randomize 200 cases for each lesion type into one of two sets (training and testing), based on lesion subtlety. Similarly, selectively randomize 200 normal cases into two sets based on breast density.
e. Place testing set in off-line storage and training cases in on-line storage.
f. On average 250 cases (2500 image -- see text for details) will be done per year for 4 years for a total of 1000 cases (10,000) images.

Task 2: Establish protocol for transmitting database. Months 1-24

a. Test protocols for different modes of transferring data between the UNC and UC (FTP, 8-mm tape, and CD). A data structure designed for portability will be provided to contain the patient text data; this data structure will be made available along with the data to the requesting sites. Use of ACR/NEMA DICOM protocol will be investigated and incorporated as an optional transfer mechanism.

Task 3: Maintain database and distribute cases Months 12-48.

a. Maintain computer, jukebox, and network connection including bug fixes and installation of vendor software updates.
b. Distribute cases via computer network and by mass storage media (tape or CD) as requested.

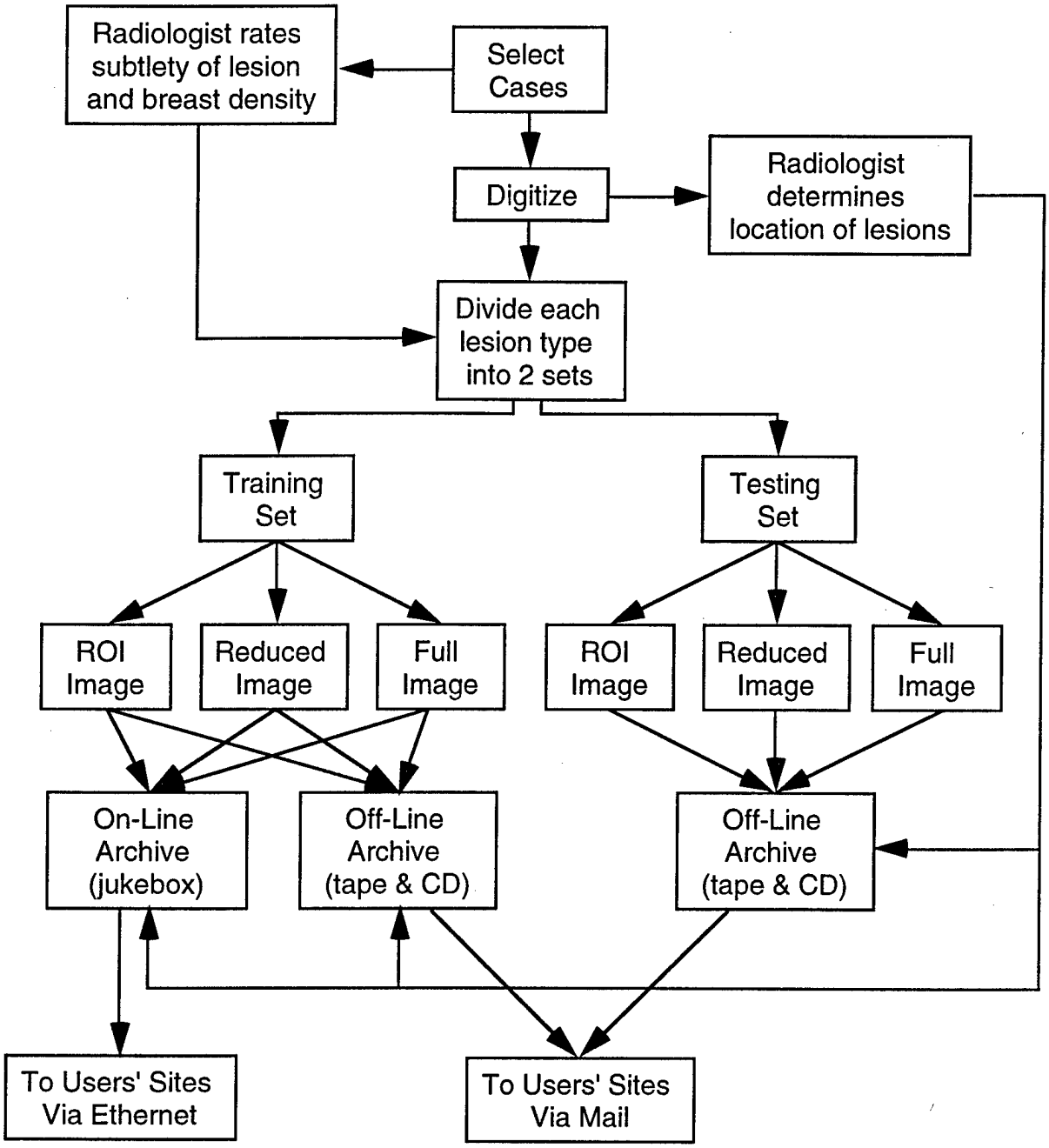


Figure 1. A flowchart of the steps required to collect, digitize, archive, and distribute the mammographic database. The 'Full Image' is the whole digitized mammogram at full resolution. The 'Reduced Image' is a minified version (reduced resolution) of the full image. The 'ROI Image' is a portion of the full image at full resolution.

6. PROGRESS TO DATE

Task 1.

We now have 448 cases digitized (see Table I, at end of report). We plan to release the first 50 cases of clustered microcalcifications immediately. We are currently verifying all the images and data before releasing these cases. We initially had planned to release 50 cases last year but we are making sure that the integrity of the data is sound. This requires input from a radiologist. Unfortunately, two radiologists in the mammography left over the past year, which slowed our effort. The department is in the process of hiring two new radiologists, which will allow us to finish verifying each case carefully. We will follow up the initial release with a release of 100 cases with masses. Further releases in batches of 100 will be done as cases accrue. Collection of cases was also been hampered by staffing problems in our mammography section. We will accelerate our case accrual to meet our goal of 1000 cases. We have identified 100 normal cases and 70 mass cases, but these need to be reviewed by a radiologist to see if they meet the requires for inclusion in the database.

All cases are archived on 4-mm or 8-mm tape.

Task 2.

We originally considered the ACR/NEMA (DICOM) image format for our database. However, the ACR/NEMA format does not have a module for mammography, and it would be an extensive project to develop one at this time. Currently, then, we are storing the images as a binary array of numbers with a simple 512-byte header. Recently, an effort has been made to establish a Digital Mammography working group in DICOM which would extend the DICOM standard to provide more specific support for digital mammographic images. One of the investigators on this project, Brad Hemminger, is a member of the committee. When the DICOM committee mammography module becomes available, it will be easy to convert our files to that format.

Task 3.

Maintenance of the database and distribution of the database are at a minimum currently. These tasks will become important shortly as cases go "on-line".

7. CONCLUSIONS

The development of a common database of mammograms for digital mammography research is underway. We are ready to release the first 50 cases of clustered microcalcifications. This will serve as a test release. We follow this initial release with an additional 50 cases of microcalcifications and 100 cases of masses. As more cases accrue, further releases will be made.

A database of mammograms would also be useful for investigators doing research in other areas of digital mammography, such as x-ray detector development, telemammography, image compression, and image processing. For example, questions such as the required spatial resolution of a digital mammogram can be answered in part by conducting observer studies using the mammograms from the database displayed at different resolutions. Furthermore, the database would provide an excellent source of cases that could be used for teaching purposes.

8. REFERENCES

1. F. Shtern, "Digital mammography and related technologies: A perspective from the National Cancer Institute," Radiology 183, 629-630 (1992).

2. R. M. Nishikawa, M. L. Giger, K. Doi, F.-F. Yin, C. J. Vyborny and R. A. Schmidt, "Effect of case selection on the performance of computer-aided detection schemes," Medical Physics 21, 265-269, (1994).

3. F. Shtern, "Panel discussion: Design of a common database for research in mammogram image analysis," Proc. SPIE 1905, 534-551 (1993).

Table I. Breakdown of cases in the database as of October 1/98.

Type of Lesion	**Pathology**	**# of Cases**
Mass	Malignant	116
Mass	Benign	75
Microcalcifications	Malignant	114
Microcalcifications	Benign	87
Asymmetric Density	Malignant	18
Asymmetric Density	Benign	4
Architectural Distortion	Malignant	19
Architectural Distortion	Benign	3
Normal		12
Total		**448**